V. SUMMARY OF CLAIMED SUBJECT MATTER

The subject matter of independent claims 1, 16, and 27 is directed to a methodology for assembling a single document image for printing from document content that spans multiple web-pages, by employing two cooperative processes. Given a starting location 110, one process analyzes a single page at a time to find candidate links 140. The links are recursively followed and those pages are analyzed. A detailed set of heuristics is used to determine what is or is not a candidate link. The links are examined for link clusters and a table of contents if found is identified. The candidate pages 120 are then fed to a document-level analyzer 150. This process compares the attributes of one page against the others and looks for a document-like structure. Using another detailed set of heuristics, the document-level analyzer 150 determines if the page should be included in the document. (see Abstract, page 20 of the specification as filed, and Figure 1)

FIG. 1

The page-level link analysis 140 is described in greater detail in Figure 2. During page-level link analysis 140, the document detection system attempts to identify links that may potentially lead to other pages within the same document. It is assumed that a well-authored multi-page document will always include progression links (links that provide some well-defined progression through the document, often indicated by the presence of some well-known contextual clue, such as a graphic or text "next" or "previous" indicator) and/or table of contents

links (clusters of links providing a path to every page or some logical subset of pages in the document) that indicate the structure of the document. These are the two categories of intra-document links that the link analysis process 140 seeks to identify. (see page 6, line 32 through page 7, line 7 of the specification as filed, and Figure 2).

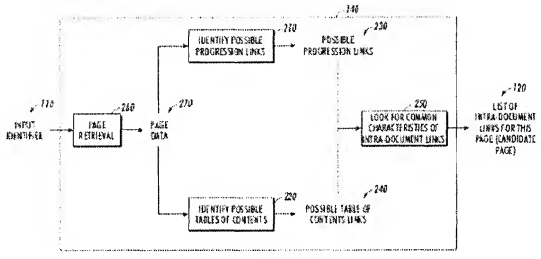

FIG. 2

The link analysis process begins with the retrieval of the actual page 270 for analysis from the page identifier 110. This is done as will be well understood by those skilled in the art, by the page retrieval process 260. The retrieved page 270 is then used as input to both the progression-link identification module 210 and the link-cluster identification module 220. In the progression-link identification module 210, possible progression links 230 are identified primarily by means of a progression indicator, which is a textual or graphical clue that suggests the nature of the progression link. Link-cluster identification module 220 examines the page data 270 to identify link clusters and thereby possible table of content type links 240. The possible progression links 230 and possible table of content links 240 are passed to module 250 for a final examination to weed out links which have properties that are not characteristic of typical intra-document links, e.g. they point to a different web server. The final result is then a

list of intra-document links 120 for the candidate page 270. (see page 7, lines 9-22 of the specification as filed, and Figure 2)

In module 250 of Figure 2 a final examination is made of all the links identified by either the progression analysis 210 or the cluster analysis 220. This module 250 identifies any hyperlinks that are significantly different in a property that is typical of intra-document links. The different link is filtered out. Thus a link to a page on a different server form all the others would be removed. (see page 10, lines 16-21 of the specification as filed, and Figure 2)

The above identified document pages are then assembled into a single document representation for subsequent printing or viewing.

There are three independent claims on appeal, Claim 1, Claim 16, and Claim 27. Claims 1, 16, and 27 are set forth below with reference numbers and citations to the specification and drawings:

1. An automated identification methodology for assembling a document representation for subsequent viewing or printing of a given hyperdocument by gathering related hyperlinked page content comprising:

performing a page-level link analysis 140 that identifies those hyperlinks on a page linking to a candidate document page 120 (p. 5, ll. 4-9; p. 6, ll. 17-20, 34-35; p. 7, ll. 1-5; Figure 1);

performing a recursive application of the page-level link analysis 140 to the linked candidate document page 120 and any further nested candidate document pages 120

thereby identified, until a collective set of identified candidate document pages 120 is assembled (p. 4, ll. 11-15; p. 6, ll. 20-27, 32-33; p. 7, ll. 9-22; p. 10, ll. 16-35; p. 11, ll. 1-14; Figures 1-2);

examining the collective set of identified candidate document pages 120 to weed out from that collective set of identified candidate document pages 120 links which have properties that are not characteristic of intra-document links, to provide a resultant set of identified candidate document pages 120 (p. 5, ll. 17-30, p. 7, ll. 24-35, p. 8, ll. 1-21; Figure 3);

grouping the content found in the resultant set of candidate document pages 120 by an automated system into a document representation 130 stored in memory by the automated system (p. 11, ll. 16-35; p. 12, ll. 1-35; p. 13, ll. 1-16; Figure 5); and,

printing, or viewing on a display by a user, the document representation 130 (p. 1, ll. 13-14, Figures 1, 5).

16. A system identification methodology for assembling a document representation 130 for subsequent viewing or printing of a given hyperlinked document comprising:

performing a page-level link analysis 140 that identifies those hyperlinks on a page linking to a candidate document page 120 further comprising a methodology of:

identifying possible progression links 210 (p. 5, ll. 4-9; p. 6, ll. 17-20, 34-35; p. 7, ll.1-3; Figure 1), and;

identifying possible table of content links 240 (p. 5, ll. 4-9; p. 6, ll. 34-35; p. 7, ll. 1-5; Figure 1);

performing a recursive application of the page-level link analysis 140 to the linked candidate document page 120 and any further nested candidate document pages 120

thereby identified, until a collective set of identified candidate document pages 120 is assembled (p. 4, ll. 11-15; p. 6, ll. 20-27, 32-33; p. 7, ll. 9-22; p. 10, ll. 16-35; p. 11, ll. 1-14; Figures 1-2);

examining the collective set of identified candidate document pages 120 to weed out links which have properties that are not characteristic of intra-document links, to provide a resultant set of identified candidate document pages 120 (p. 5, ll. 17-30, p. 7, ll. 24-35, p. 8, ll. 1-21; Figure 3);

grouping the resultant set of candidate document pages 120 by an automated system into a document representation 130 stored in memory by the automated system (p. 11, ll. 16-35; p. 12, ll. 1-35; p. 13, ll. 1-16; Figure 5); and,

printing, viewing on a display by a user, the document representation 130 (p. 1, ll. 13-14, Figures 1, 5).

27. A system identification methodology for assembling a document representation 130 for later viewing or printing of a given hyperlinked document comprising:

performing a page-level link analysis 140 that identifies those hyperlinks on a page linking to a candidate document page 120 further comprising a methodology of:

identifying possible progression links 210 (p. 5, ll. 4-9; p. 6, ll. 17-20, 34-35; p. 7, ll.1-3; Figure 1);

identifying possible table of content links 240 (p. 5, ll. 4-9; p. 6, ll. 34-35; p. 7, ll. 1-5; Figure 1), and;

examining the possible progression links and the possible table of content links for common characteristics (p. 6, ll. 17-27; p. 9, ll. 26-30; Figure 1);

performing a recursive application of the page-level link analysis 140 to the linked candidate document page 120 and any further nested candidate document pages 120 thereby identified, until a collective set of identified candidate document pages 120 is assembled (p. 4, ll. 11-15; p. 6, ll. 20-27, 32-33; p. 7, ll. 9-22; p. 10, ll. 16-35; p. 11, ll. 1-14; Fig 1-2);

examining the collective set of identified candidate document pages 120 to weed out links which have properties that are not characteristic of intra-document links, to provide a resultant set of identified candidate document pages 120 (p. 5, ll. 17-30, p. 7, ll. 24-35, p. 8, ll. 1-21; Fig. 3);

grouping the resultant set of candidate document pages 120 by an automated system into a document representation 130 stored in memory by the automated system (p. 11, ll. 16-35; p. 12, ll. 1-35; p. 13, ll. 1-16; Figure 5); and,

printing, or viewing on a display by a user the document representation 130 (p. 1, ll. 13-14, Figures 1, 5).